Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Winter 2019/2020

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Mapping and Understanding the Evolution of Language

Our natural language is constantly evolving. The words we use change over time, but also their meaning or the context in which we use them. One word can even mean different things, for example, “apple” is a fruit or a company. In the research area of Natural Language Processing (NLP), there are already models that try to analyze evolving language or that automatically identify words with multiple senses. However, doing both at the same time or successfully using such models in applications for automated text processing is only now moving into the focus of current research.

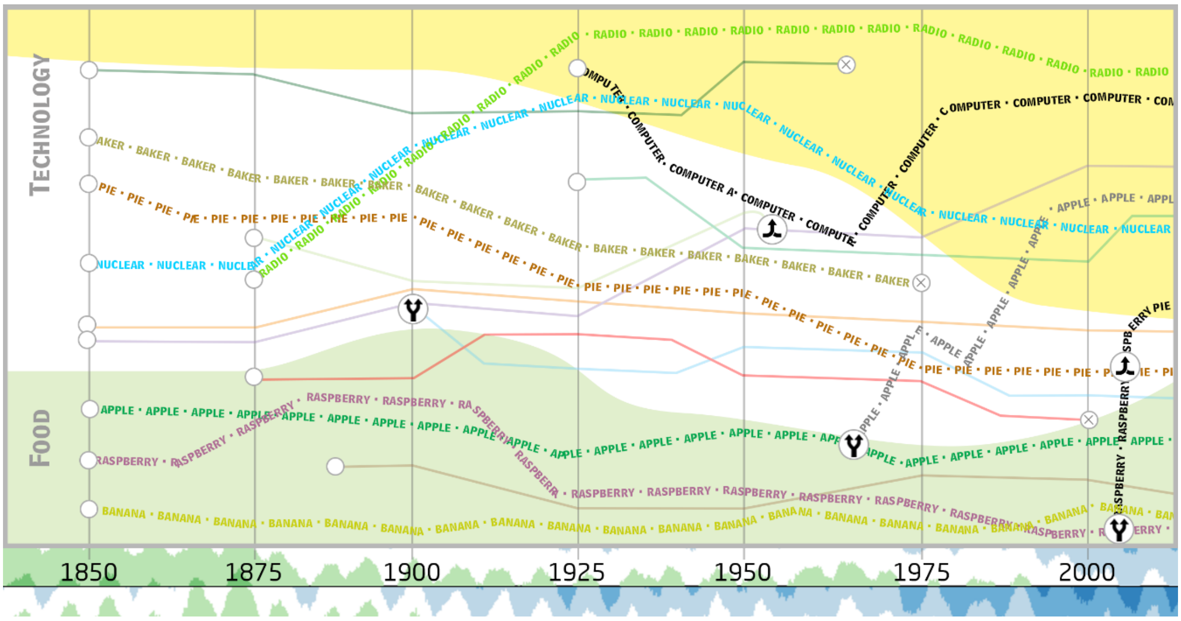

Figure 1: A vision mockup of our language model visualization

In this master’s project, we will develop a novel approach that represents the usage of words over time as a graph. Using carefully crafted restrictions to the two-dimensional layout when drawing that graph, we will have fine-grained control to understand the resulting language models. Although this is mainly a computational problem, we will focus on visualization subsets of the model to explore the feasibility and quality of such an approach.

The visualization will show the graph structure of selected words and their context as used in the underlying text corpus. In a large timeline, where each line represents a word, it will be possible to see how they appear, split into different meanings, or disappear again. The closeness of lines in the resulting chart conveys semantic similarity. The figure above shows a sketch of an interactive visualization that we aim to develop. This will help to investigate the feasibility and quality of this approach and provide a better way for qualitative evaluation of dynamic language models.

Students will train word embeddings on documents from different time slices. In the visualization, each time slice is a vertical axis on which words are placed based on their similarity in the embedding. After connecting the words across time slices, we the weighted edges as force constraints. An iterative algorithm will try to optimize network layout by reordering nodes (i.e. words on the axes) or splitting them. For the visualization, we only draw selected words and their neighbors, and use additional meta-data to make the timeline more interesting, for example, topic zones, frequencies as sizes, or color for highlights.

Contact

We offer this project for up to five students who will be supervised closely by Prof. Felix Naumann and Tim Repke (tim.repke@hpi.de, F-2.07).
If you have any questions, please do not hesitate to contact us.

References

- Kutuzov, Øvrelid, Szymanski, Velldal. *Diachronic word embeddings and semantic shifts: a survey*. ICCL 2018
- Rule, Cointet, Bearman. *Lexical shifts, substantive changes, and continuity in State of the Union Discourse, 1790-2014.* PNAS 2015
- Mitra, Mitra, Riedl, Biemann, Mukherjee, Goyal. That’s sick dude!: Automatic identification of word sense change across different timescales. arXiv 2014
- Bamler, Mandth. *Dynamic Word Embeddings.* PMLR 2017
- Devlin, Chang, Lee, Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv 2019
- Nguyen, Nguyen, Vung, Dang*. FinanViz: Visualizing Emerging Topics in Financial News.* BigData 2018
- Liu, Cui, Wu, Liu. *A survey on information visualization: recent advances and challenges.* VisComput 2014
- Brath, Banissi. *Microtext Line Charts.* IV 2017
- https://hpi.de/naumann/projects/rdbms-genealogy.html